

Supervised Attribute Relevance Determination for Protein Identification in Stress Experiments

M. Strickert^a, K. Witzel^a, H.-P. Mock^a, F.-M. Schleif^b, and T. Villmann^b

^a Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben,
{stricker,witzel,mock}@ipk-gatersleben.de

^b University of Leipzig, Medical Department,
{schleif,villmann}@informatik.uni-leipzig.de

Abstract. Biomedical high-throughput measuring devices, used in many stages of data acquisition related to systems biology research, provide access to huge-dimensional data descriptions that create particular challenges in data analysis. One major issue to meet is that data spaces do often exhibit a much higher dimensionality than there are experiments available. For comparative analysis of labeled experimental data we propose an efficient algorithm with parallel attribute weighting for reducing the influence of unnecessary data attributes. The method is based on a mathematical cost function by means of which the parameters of adaptive data metrics are determined in an asymptotically optimum manner. The approach is quite generic with respect to the choice of the underlying data metric. Very insightful results are obtained for the identification of preferentially expressed proteins in a current 2D gel-plot study.

Key words: Feature rating, adaptive metrics for labeled data.

1 Introduction

Massive data sets with a high number of samples and/or attributes constitute the major basis for systems biology research. Particularly, high-throughput biomedical devices like gene expression arrays or tandem spectrometers for coupled chromatogram and mass detection generate thousands of data points in parallel. Feature selection techniques do substantially help reducing data volumes and thus to lower the complexity of problems related to pathway inference and modeling of regulatory networks. We propose a new method for parallel assessment of the discriminative power of data attributes in annotated experimental data, labeled by class information, such as the type of treatment.

Traditionally, hypotheses are formulated about attributes, such as genes, proteins or masses, differentiating between the group of treated experiments and the non-treated control group; statistical tests like t-test, U-test or F-test are conducted for taking the decision whether to accept or reject the suppositions at a certain level of significance. However, there is not only well-argued criticism about such find of statistical testing [3], but additionally, the problem of dealing simultaneously with more than two experimental conditions remains.

In the area of machine learning the detection of interesting data attributes is known as the variable subset selection problem, asking which combination of data attributes does explain specific class assignments [1]. Filter methods make use of statistical properties such as variance or information gain criteria for rating the data attributes. Complementary to that, wrapper methods apply an external feature selection scheme, such as random sampling, and rate the explanatory power of the reduced data set by passing it to a separate classifier for which the classification success should be maximum. Filters are usually fast, because they do not require construction of attribute probing classifiers, but their truth is strictly dependent on the chosen filter criterion. In contrast to that, exhaustive attribute probing is known to be NP-hard and is thus hardly applicable in large wrapper scenarios. The class of so called embedded methods is located between the computing complexity of filters and wrappers. Embedded methods, like the one proposed in the following, perform implicit attribute selection already during model construction.

We propose a method for supervised attribute relevance determination using cross comparisons (SARDUX) which has several appealing properties:

1. The algorithm is very simple, leading to easily communicable results.
2. The runtime is $\mathcal{O}(n^2)$, n denoting the number of data samples; all pairs of experiments are compared, which is the preferred style of biomedical researchers.
3. Its mathematical cost function optimization leads to high reproducibility.
4. Attributes are rated on a real value scale rather than switched on or off in an attribute selection scenario; rating interpretation is very intuitive.
5. Custom similarity measures or data metrics can be used if they can be formulated in terms of an 'adaptive' version as discussed below.

The basic idea of the new method is taken from generalized relevance learning vector quantization (GRLVQ), a centroid-based classification system with cost function driven updates of centroid locations and attribute relevance factors [2]. In the present case data quantization into a centroid representation is not desired in order to maintain all information in the usually small number of available experiments. Still, we make use of data-driven metric adaptation which is accomplished by cost function optimization, a general approach that is also shared by other researchers [4].

2 Method

Simultaneous attribute characterization can be considered as a special case of application-driven metric adaptation. Here, class information is considered to define subsets of data that should be most similar within the respective class and maximum dissimilar between different classes.

The following cost function is optimized over pairs of all n data items:

$$s = \frac{\sum_{i=1}^n \sum_{j=1}^n d_{\lambda}(\mathbf{x}^i, \mathbf{x}^j) \cdot \delta_{ij}}{\sum_{i=1}^n \sum_{j=1}^n d_{\lambda}(\mathbf{x}^i, \mathbf{x}^j) \cdot (1 - \delta_{ij})} = \frac{d_C}{d_D} = \min . \quad (1)$$

Distances and, likewise, dissimilarity measures $d_{\lambda}^{ij} = d_{\lambda}(\mathbf{x}^i, \mathbf{x}^j)$ between data vectors \mathbf{x}^i and \mathbf{x}^j do depend on our adaptive parameters $\lambda = (\lambda_k)_{1\dots m}$ of interest. The Kronecker symbols δ_{ij} indicate identity of class memberships, i.e. $\delta_{ij} = 1$ if the class of data vector \mathbf{x}^i equals the class of vector \mathbf{x}^j , $\delta_{ij} = 0$ in the case of disagreement. The stress function s thus penalizes parameters λ_k that would lead to a contraction of data vectors belonging to different classes. Conversely, by minimizing s , large denominators d_D are preferred in Eqn. 1 that express a large distance between data of different classes, and small numerators d_C are targeted that reflect small distance of data belonging to the same class.

Cost function s is optimized iteratively by adapting each component k of the parameter vector λ in small steps $\gamma < 1$ into the direction of steepest gradient according to the derivative of the cost function, obtained by the chain rule:

$$\lambda_k \leftarrow \lambda_k - \gamma \cdot \frac{\partial s}{\partial \lambda_k} \quad \rightarrow \quad \frac{\partial s}{\partial \lambda_k} = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial s}{\partial d_{\lambda}^{ij}} \cdot \frac{\partial d_{\lambda}^{ij}}{\partial \lambda_k} = \min \quad (2)$$

By applying the quotient rule to the fraction $s = d_C/d_D$ this yields

$$\frac{\partial s}{\partial d_{\lambda}^{ij}} = \frac{\delta_{ij} \cdot d_D}{d_D^2} + \frac{(\delta_{ij} - 1) \cdot d_C}{d_D^2}. \quad (3)$$

If the classes of vectors \mathbf{x}^i and \mathbf{x}^j are identical only the left summand is evaluated, otherwise the right one takes over. For m -dimensional data comparison we chose the parametric Euclidean metric with its intuitive weighting of data attributes:

$$d_{\lambda}^{ij} = \sqrt{\sum_{k=1}^m \lambda_k \cdot (x_k^i - x_k^j)^2} \quad \rightarrow \quad \frac{\partial d_{\lambda}^{ij}}{\partial \lambda_k} = \frac{(x_k^i - x_k^j)^2}{2 \cdot d_{\lambda}^{ij}}. \quad (4)$$

The parameters λ_k control the influence of the k -th data attribute under the optimization condition. They are initialized by unity $\lambda_k = 1$ which yields ordinary Euclidean distance. Then, they are iteratively adapted according to Eqn. 2 for minimizing the cost function. After each update step, the parameter vector is normalized to $\sum_{k=1}^m \lambda_k = m$. The update loop ends when changes in s subside. Large final values of λ_k indicate attributes that increase class separation, low values reflect either noise or suppressed intra-class differences. Note that even small differences get magnified and detected if they are meaningful.

3 Applications

Synthetic data. For illustration we use a rather difficult 10-dimensional data set with 180 points and 3 overlapping multimodal classes as known from [2]. The first two attributes are the generating dimensions producing Gaussian data clouds in $[0; 1]^2$. Attributes 3–6 are copies of the first one, contaminated by Gaussian noise of variances 0.05, 0.1, 0.2, and 0.5, respectively; attributes 7–8 contain pure uniform noise in $[-0.5; 0.5]$ and $[-0.2; 0.2]$, 9–10 reflect Gaussian noise with variances 0.5 and 0.2. Figure 1, top row, shows the principal component

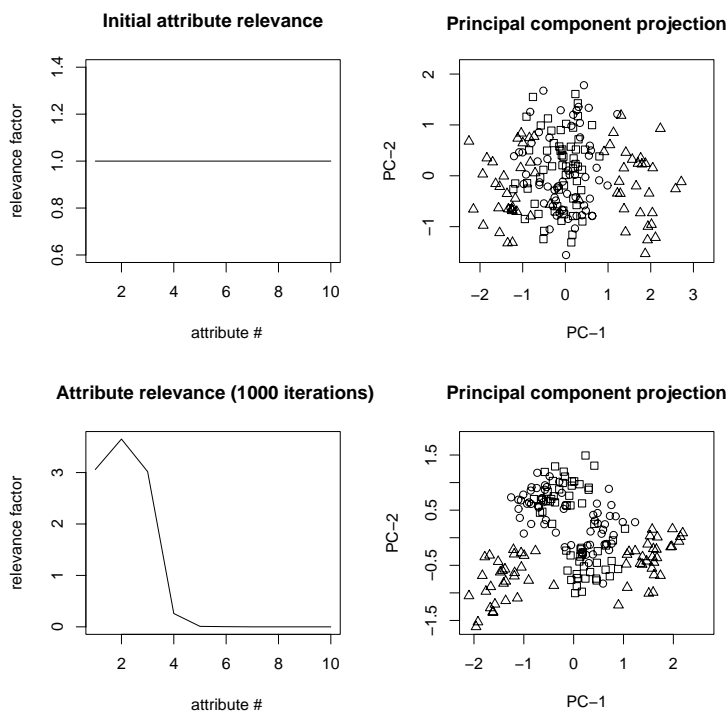


Fig. 1. Synthetic 10-dimensional 3-class data set before (top) and after (bottom) application of the obtained metric parameters (left) to the data set (right, shown as projection to first two principal components).

projection of the data at initialization time with $\lambda_k = 1$; the bottom row contains the SARDUX results after 1000 updates of the metric. As main results, the first two generating attributes are clearly recovered and the class separation of the projected λ -rescaled data becomes apparent; the low-noise attributes 3 and 4 are maintained, because they do not much affect separation.

Protein gel data. The nutrition uptake and the effects of environmental stress in barley roots are of high interest for breeding companies. Salt stress experiments were conducted at 100mM concentration level with the tolerant Morex line and the sensitive Steptoe line. Protein abundances were obtained by electrophoresis in 2D-gels, scanned and spotted by commercial software. Spots of six gels, three technical replica per line, were quantile normalized and subject to SARDUX analysis. Figure 2 shows two exemplary spots pointed out by metric adaptation after 1000 iterations. Spot 654, identified via additional mass spectrometry as glutathione transferase, is a very obvious separator. However, spot 452 ranked at λ_k -order 14, belonging to the initiation factor 2 subunit family, is a very interesting candidate that would have been missed if the standard two-fold change criterion between spot volumes would have been applied.

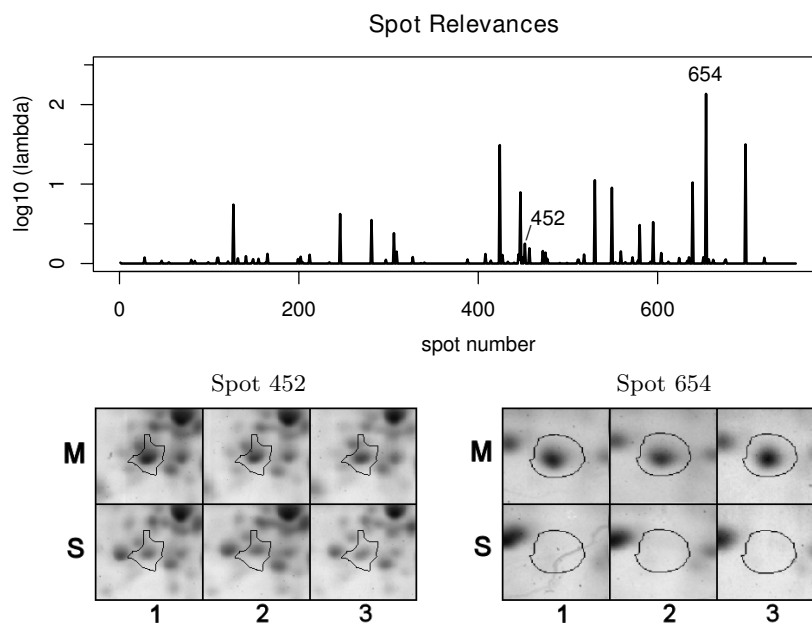


Fig. 2. Spot rating in 2D gel plots. Top: $\log\text{-}\lambda$ profile with $\lambda_k \geq 1$. Bottom: 2D-gels of both replica series zoomed to spot numbers indicated above (M=Morex, S=Steptoe).

4 Conclusions

The presented SARDUX method is proposed for identification of class-separating metric parameters by using a mathematically sound framework. For Euclidean distance, intuitive attribute ratings are obtained that are significant for synthetic and real-world data. Further work will focus on the biological surplus value of the results, especially, in comparison to other methods. Moreover, outcomes from alternative data measures, like parametric Pearson correlation, will be studied. This work is supported by the Ministry of Culture of Saxony-Anhalt, grant XP3624HP/0606T.

References

1. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7–8):1157–1182, 2003.
2. B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.
3. D. Johnson. The insignificance of statistical significance testing. *Journal of Wildlife Management*, 63(3):763–772, 1999.
4. S. Kaski. Learning metrics for exploratory data analysis. In D. Miller, T. Adali, J. Larsen, M. V. Hulle, and S. Douglas, editors, *Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pages 53–62. IEEE Computer Society, 2001.